DOCUMENT RESUME

ED 281 360 FL 016 633

AUTHOR Brutten, Sheila R.; And Others
TITLE Identifying Discrepant Item Response Patterns (Item Bias) in a Measure of ESL Subjects' Control of the (D) and (Z) Morphemes.
PUB DATE [87]
NOTE 33p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Articulation (Speech); *Chinese; Comparative Analysis; Correlation; Difficulty Level; *English (Second Language); *Item Analysis; Morphemes; *Morphophonemics; Phonology; Pronunciation; Second Language Learning; *Spanish Speaking; Statistical Analysis; Test Bias; *Test Items; Uncommonly Taught Languages

ABSTRACT

In a study of the detection of test item bias, Chinese speakers and Spanish speakers were administered a measure of pronunciation accuracy for the (d) and (z) morphemes. Six indices were calculated for each item: item difficulty, point biserial correlation coefficient, item variance, chi-square, Rasch logit of difficulty, and the average standardized difference between the expected probability of a correct response and the observed item response. Twelve different items were flagged as biased by one or more indices. Five items were flagged for both language samples. The classical item indices, item difficulty and the point biserial correlation coefficient, flagged more items than the remaining indices combined. This finding can be explained by the fact that both samples exhibited similar levels of performance on the pronunciation task. Analyses of variance indicated a significant effect for the choice of allomorph and the stem final place of articulation for the conditioning of the allomorphs. (Author/MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED281360

Identifying Discrepant Item Response Patterns (Item Bias) in a Measure of ESL Subjects' Control of the {D} and {Z} Morphemes

Sheila R. Brutten, John T. Mouw, and Kyle Perkins

Southern Illinois University

Abstract

A sample of Chinese speakers and a sample of Spanish speakers were randomly selected for this study. A measure of their pronunciation accuracy of the {D} and {Z} morphemes was taken. Six indices were calculated for each item in a study of item bias: (1) item difficulty, (2) point biserial correlation coefficient, (3) item variance, (4) Chi-square, (5) Rasch logit of difficulty, and (6) the average standardized difference between the expected probability of a correct response and the observed item response.

Twelve different items were flagged as biased items by one or more indices. Five items were flagged for both language samples. The classical item indices, item difficulty and the point biserial correlation coefficient, flagged more items than the remaining indices combined. This finding can be explained by the fact that both samples exhibited similar levels of performance on the pronunciation task. Analyses of variance indicated a significant effect for the choice of allomorph and the stem final place of articulation for the conditioning of the allomorphs.

FL016633

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

K. Perkins

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Introduction

An important consideration in test development and test use is the identification and elimination of biased items. In this context an item is considered to be biased if persons of equal ability but from different cultural, ethnic, gender or language groups do not have equal chance or opportunity of answering the item successfully.

Generally two methods of detecting item bias have been employed: (1) an item-by-group interaction (e.g., Angoff and Ford, 1973) and (2) determining the probability of a correct response for different groups depending on their ability level (e.g., Shepard, Camilli, and Williams, 1985). The item-by-group interaction approach attempts to identify items that are relatively more difficult for members of a particular group when the total test is used as the overall measure of ability. The second approach entails the calculation of the difference between the probability of a correct response and the observed item response for a particular group and noting whether the differences of that particular group deviate significantly from the majority or the criterion group.

The present study compares six methods of identifying discrepant item response patterns (item bias) in a measure of ESL subjects' control of the {D} and {Z} morphemes.

The {D} morpheme is found in the past tense endings for regular verbs, the past participle or third principal part of a verb, and the -ed past participle adjective. Although these endings are spelled with ed (d, ied), there are three different pronunciations (allomorphs): /t/, /d/, and /id/. The -ed ending is pronounced /t/ after all voiceless consonants except /t/; /d/ after all voiced consonants except /d/, and after all vowel sounds; and /id/, as a separate syllable, after /t,d/.

The {z} morpheme is found in the plural of regular nouns, the possessive form of nouns, and in the third person singular present tense marker for verbs. Although these endings are spelled with s ('s, s', es, ies), there are three different pronunciations (allomorphs): /s/, /z/, /ɨz/. The -s ending is pronounced /s/ after all voiceless consonants except /s/ and /š/; /z/ after all voiced consonants except /z/ and /ž/ and after all vowel sounds; and /ɨz/, as a separate syllable, after /s, š, z, ž/. The same conditioning applies when the singular present tense form of the verb be (is) is contracted with a noun.

METHOD

Subjects

This study was conducted at the Center for English as a Second Language at Southern Illinois University. Twenty four Chinese speakers from level four and four Chinese speakers from level three were randomly selected to form the Chinese sample for this study. Twenty four Spanish speakers from level four and four Spanish speakers from level three were randomly selected to form the Spanish sample.

At the Center for English as a Second Language, placement of students into the four full-time proficiency levels (1, 2, 3, and 4) is determined by the results of an Institutional TOEFL test which is administered to new students at the beginning of each term. Students who score 349 or lower are placed into level 1; students who score between 350 and 399 are placed into level 2; those who score between 400 and 449 are placed into level 3; and those who score between 450 and 499

are placed into level 4. If students score between 500 and 524, they are given additional testing (a composition and an oral interview) to determine whether they are proficient enough for full- or part-time university study or whether they need additional full- or part-time English study at CESL.

Instrumentation

We administered a pronunciation test in the language laboratory in order to elicit data. The test appears as Appendix A. The underlined word in each sentence is the stimulus for the allomorph elicitation.

(Put Appendix A here)

Appendix B contains a list of the morpheme, the stem final sound, and the allomorph for each stimulus sentence.

(Put Appendix B here)

PROCEDURES

We calculated six indices for each item: (1) item difficulty, (2) point biserial correlation coefficient, (3) item variance, (4) Chi-square, (5) Rasch logit of difficulty, and (6) the average standardized difference between the expected probability of a correct response and the observed item response.

Item difficulty is calculated as the proportion of correct responses. We flagged an item as being biased if its proportion of correct answers was less than .33. If one were interested in positive bias, one would flag items with difficulty indices higher than .67, but we are only concerned here with negative bias. The raw difficulty (p value) of each item was calculated separately for each group.

The point biserial correlation coefficient is a common procedure for computing item discriminability. This procedure involves an item-total test correlation for each item in which the binary item scores are correlated with the continuous total test scores. If an item had a point biserial correlation coefficient of less than .25 and also had a p value of less than .33, we flagged it as a biased item. Obviously items with very high p values will also have low point biserial correlation coefficients because they do not discriminate very well. The point biserial correlation coefficient of each item was calculated separately for each group.

Item variance is calculated by multiplying the proportion correct by the proportion incorrect for each item. We set .10 as the threshold. Items with a variance of less than .10 provide very little information function about the item. If an item had a variance of less than .10 and a p value of less than .33, it was flagged as being biased. If an item has a high p value, it will by necessity also have a low variance. The item variance of each item was calculated separately for each group.

We calculated a one-way Chi-square for each item. We used the marginal frequencies (item responses for both groups, i.e., column totals and total scores for each group, i.e., row totals) to determine the expected cell frequencies for each item using the following formula:

$$E_{ij} = \frac{n_i n_j}{N}$$

Since our design was one-way and had one d.f., we corrected the estimates for attenuation by subtracting .5 from the observed value if it was larger than the expected value and by adding .5 to the observed value if it was less than the expected value. If an item had a

Chi-square greater than 3.841 ($p < .05$), we flagged it as being biased. The Chi-square of each item was calculated separately for each group.

Rasch one-parameter item response theory logits of difficulty corrected for the sample spread expansion factor were calculated for each item using the PROX procedure (Wright and Stone, 1979). Difficulty calibrations were calculated from the responses of each group separately and then from the analysis of the combined responses of the two groups. Rasch negative logits of difficulty indicate comparative ease of the items, and positive logits of difficulty indicate comparative difficulty of the items. The Rasch model requires that there must be at least one measurable person for each item; therefore, items for which all responses were correct or all incorrect are removed from the analysis. All the Chinese subjects answered item number 11 correctly, and that item was removed from the Rasch analysis for both groups.

No formal statistical decision theory exists for interpreting Rasch logits of difficulty as bias indices. Shepard et al. (1981) recommended that biased items should appear as outliers when they are compared to other items intended to measure the same trait. To that end the Rasch logits of difficulty calculated from the analysis of the combined responses of the two groups and the Rasch logits calculated from the Chinese group were submitted to a least-squares regression procedure, and a 95 percent confidence interval was constructed around the regression line. Biased items were items whose item coordinates fell outside the 95 percent confidence interval. The same procedure was repeated for the Spanish group. Figure 1 gives an illustration of the outlier concept.

(Put Figure 1 here)

Following Linn and Harnish (1981), we computed the average standardized difference (Z_{ig}) between the expected probability of a correct response (P'_{ij}) and the observed item response (U_{ij}) for each item using the following formula:

$$Z_{ig} = \frac{1}{n} \sum_{j \in g} \frac{U_{ij} - P_{ij}}{\sqrt{P_{ij}\,(1-P_{ij})}}$$

where $U_{ij} = 1$ if person j answers item i correctly, and $U_{ij} = 0$ if person j answers item i incorrectly.

For this analysis we used the Rasch logits of difficulty from the previous analysis. The Rasch item difficulty estimates and person ability estimates were used to determine the probability of a correct response (P_{ij}) by means of the conversion table in Wright and Stone (1979).

(Put Table 1 here)

The average standardized difference between the expected probability of a correct response and the observed item response was calculated for each item separately for each group and then from the analysis of the combined responses of the two groups. These data were submitted to a least-squares regression procedure, and a 95 percent confidence interval was constructed around the regression line. Biased items were those whose item coordinates fell outside the 95 percent confidence interval.

RESULTS

The means and standard deviations for the two language groups are listed below.

Chinese (n=28)	Spanish (n=28)
$\bar{x}$ = 24.89	$\bar{x}$ = 25.86
s = 7.39	s = 6.77

The internal consistency reliability estimated by the Kuder-Richardson Formula 20 procedure was .857.

The Rasch model (and other item response theory models) assumes unidimensionality of items, i.e., the underlying construct purportedly measured by a test can be described in a one-dimensional latent space (Lord and Novick, 1968; Reckase, 1979; Whitely and Davis, 1974). We used Guttman implicational scaling techniques to determine whether our data were truly scalable and unidimensional. For a data set to be truly scalable and to exhibit unidimensionality, the obtained coefficient of scalability "should be well above .6" (Nie, Hull, Jenkins, Steinbrenner and Bent 1975:533).

The Guttman coefficients for our data set were the following:

Coefficient of reproducibility	.75
Minimum marginal reproducibility	.36
Percent improvement	.39
Coefficient of scalability	.61

Obviously the coefficient of scalability for this data set is not well above .6, but we believe that we can safely apply the Rasch model to this data set because of some recent research on the assumption of

unidimensionality and item response theory models. For example, Henning, Hudson, and Turner (1985) have argued that the Rasch model is sufficiently robust with respect to unidimensionality to permit its application to test data representing different item domains and to data elicited from subjects who come from different cultural, educational, and linguistic backgrounds. Another case in point is Dorans and Kingston (1985). In a study of item response theory equating, Dorans and Kingston have demonstrated that "although violations of unidimensionality have an impact on equating, the effect may not be substantial" (p. 249).

(Put Tables 2 and 3 here)

Tables 2 and 3 present the item data from the six analyses for both language groups. Below are two summary tables.

Chinese

Item Number	Morpheme	Item Difficulty	Point Biserial	Item Variance	Chi-Square	Rasch Outlier	Z_{ij} Outlier
3	Z	x	x				
12	D	x					
15	Z	x					
22	Z	x	x				
25	Z		x				
26	D		x				
28	D						x
29	Z	x	x				
33	D					x	
37	D	x	x				

Spanish

Item Number	Morpheme	Item Difficulty	Point Biserial	Item Variance	Chi-Square	Rasch Outlier	Z_{ij} Outlier
1	D						x
3	Z	x					
12	D	x					
18	D	x					
22	Z	x					
29	Z	x					
37	D	x					

For the Chinese sample ten items were flagged as biased by one or more methods; seven items were flagged for the Spanish sample. Five items (3, 12, 22, 29, and 37) were flagged for both samples.

There were no items having a variance of less than .10 and a p value of less than .33. One can conclude that item variance was not a very useful statistic for detecting item bias for this scale. No items were flagged by the Chi-square analysis which indicates that these two language groups experienced difficulty with many of the same items. These results might have been different if different language groups had been employed. We believe that the same interpretation applies to the two outlier analyses. Only three items were flagged by the two outlier analyses. One of these items (Chinese, 33) has a negative logit of difficulty (-1.02). Comparatively speaking, this item must have been easier than the other easy items for the Chinese group according to the outlier procedure.

Thus far in this paper we have been concerned with the item responses from two language samples and various item statistics generated from those responses. These item responses describe a level of performance at a given point in time for the two language samples. We wanted to know why the two samples obtained the item scores they did, and to answer that question, we ran two-way analyses of variance on the item responses. We were interested in the effects of (a) allomorph choice (b) the place of articulation of the stem final sound, and (c) the manner of articulation of the stem final sound. As we pointed out in a previous section of the paper, the stem final sound determines the choice of the following allomorph. Since all the stem final sounds in our stimuli were consonants, we tested for the effects of manner and

place of articulation because consonants are grouped according to how the sounds are produced, i.e., stops, fricatives, affricates, nasals, and liquids, and according to the place of articulation, i.e., bilabials, labiodentals, interdentals, alveolars, palatals, and velars.

(Put Figures 2, 3, and 4 here)

Figures 2, 3, and 4 present distributions by allomorph choice, by the manner of articulation of the stem final sound, and by the place of articulation of the stem final sound. Figure 2 shows that the /s/ allomorph is comparatively easy for both groups while the /t/ and /ɨz/ allomorphs are comparatively more difficult. Figure 3 indicates that if a word stem ends with an affricate, these subjects are likely to have difficulty with the allomorph. Figure 4 suggests that these subjects will likely have difficulty producing the correct allomorphs for words ending in interdentals.

The results of the analyses of variance are the following:

Allomorphs

Effect of language groups	$F(1,70) = 0.20$, n.s.
Effect of allomorphs	$F(5,70) = 3.61$, $p < .01$
Interaction	$F(5,70) = 1.43$, n.s.

Stem final manner of articulation

Effect of language groups	$F(1,72) = 0.24$, n.s.
Effect of manners of articulation	$F(4,72) = 1.92$, n.s.
Interaction	$F(4,72) = 0.45$, n.s.

Stem final places of articulation

Effect of language groups	$F(1,70) = 0.27$, n.s.
Effect of places of articulation	$F(5,70) = 3.89$, $p < .01$
Interaction	$F(5,70) = 0.10$, n.s.

For this data set our subjects' control of the {D} and {Z} morphemes is heavily dependent on the allomorph in the stimulus and the place of articulation of the stem final sound. No significant

differences are attributable to the language groups which is in concert with the paucity of items flagged by the outlier analyses, with no items having been flagged by the Chi-square, and with the fact that the averages of the two language groups were virtually the same.

We conclude this discussion of item bias indices on a cautionary note. Bias indices can be wrong for at least two reasons. First, they may indicate the presence of some extraneous, confounding factor which is not related to the purported construct being measured. Nonrandom, extraneous variance is a factor that must be reckoned with in all types of measurement studies. Second, bias indices may be wrong due to bias in sampling or other sampling peculiarities. Again, this is a factor that must be controlled for in all measurement studies either by the design of the study itself or by statistical procedures.

In conclusion, our purpose in conducting this study was to compare six methods of identifying discrepant item response patterns (item bias) in a measure of ESL subjects' control of the {D} and {Z} morphemes. The results are that the traditional item statistics (item difficulty and point biserial correlation) flagged more items than item variance, the Chi-square, and the outlier techniques derived from Rasch item difficulty indices. Both language samples tended to experience difficulty with the same items which is the reason why the Chi-square and the outlier procedures did not flag as many items as item difficulty and the point biserial correlation. The choice of the allomorph and the place of articulation of the stem final conditioning environment had significant effects on the item responses.

References

Angoff, W. H., and S. F. Ford. 1973. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement 10: 95-105.

Dorans, N. J., and N. M. Kingston. 1985. The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. Journal of Educational Measurement 22: 249-262.

Henning, G. H., T. Hudson, and J. Turner. 1985. Item response theory and the assumption of unidimensionality for language tests. Language Testing. Forthcoming.

Linn, R. L., and D. L. Harnisch. 1981. Interactions between item content and group membership on achievement test items. Journal of Educational Measurement 18: 109-118.

Lord, F. M., and M. R. Novick. 1968. Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Nie, N. H., C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Bent. 1975. Statistical package for the social sciences. Second edition. New York: McGraw-Hill.

Reckase, M. D. 1979. Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics 4:207-230.

Shepard, L. A., G. Camilli, and M. Averill. 1981. Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics 6: 317-375.

Shepard, L. A., G. Camilli, and D. M. Williams. 1985. Validity of approximation techniques for detecting item bias. Journal of Educational Measurement 22: 77-105.

Whitely, S. E., and R. V. Davis. 1974. The nature of objectivity with the Rasch model. Journal of Educational Measurement 11: 163-178.

Wright, B. D., and M. H. Stone. 1979. Best test design: Rasch measurement. Chicago: MESA Press.

Appendix A
Stimuli Sentences

INSTRUCTIONS:

This is a pronunciation test. We want to see how well you can pronounce English words. Please state your full name, your class, and your native language. Then, read each sentence on the test paper in front of you aloud.

1. He breathed deeply.
2. His name's familiar.
3. Jess's mother is here.
4. He feared snakes.
5. His worst fears came true.
6. Bob's team lost the game.
7. She smiled at me.
8. The plant bloomed.
9. The cook mashes potatoes
10. A fisherman braves the storm
11. Jeff's car is broken.
12. The player became fatigued.
13. He gave me two prizes.
14. I liked the play.
15. Two dudes tried to ride a horse.
16. Mary digs in her garden every day.
17. You surprised me.
18. She wedged the door.
19. James robbed the bank.
20. We stuffed envelopes.
21. I fanned myself.
22. The minister accepted two tithes.
23. The attendant seated us.
24. He banged the drum.
25. We saw two car crashes.
26. We bunched the celery.
27. He bangs the drum.
28. He seeded the lawn.
29. The actress rouges her lips.
30. We visited two towns.

31. The car backs up to the line.
32. Her lips are red.
33. She mashed the potatoes.
34. The farmer hauls his crop to market.
35. My cat caught two moths.
36. We passed the car.
37. The soup frothed.
38. He tapped his pipe.
39. The teacher applied for many grants.
40. Tom saved the book.
41. The azure's colorful.

Appendix B

Morpheme, Stem Final Sound, and Allomorph in Each Sentence

Item Number	Morpheme	Stem Final Sound	Allomorph
1.	D	ə	d
2.	Z	m	z
3.	Z	s	ɨz
4.	D	r	d
5.	Z	r	z
6.	Z	b	z
7.	D	l	d
8.	D	r̥	d̥
9.	Z	š	ɨz
10.	Z	v	z
11.	Z	f	s
12.	D	g	d
13.	Z	z	ɨz
14.	D	k	t
15.	Z	d	z
16.	Z	g	z
17.	D	z	d
18.	D	j	d
19.	D	b	d
20.	D	f	t
21.	D	n	d
22.	Z	ə	z
23.	D	t	ɨd
24.	D	ŋ	d
25.	Z	š	ɨz
26.	D	č	t
27.	Z	ŋ	z
28.	D	d	ɨd
29.	Z	ž	ɨz
30.	D	t	ɨd
31.	Z	k	s

Item Number	Morpheme	Stem Final Sound	Allomorph
32.	Z	p	s
33.	D	š	t
34.	Z	l	z
35.	Z	θ	s
36.	D	s	t
37.	D	θ	t
38.	D	p	t
39.	Z	t	s
40.	D	v	d
41.	Z	r	z

Figure 1

Regression Line with Outliers Example

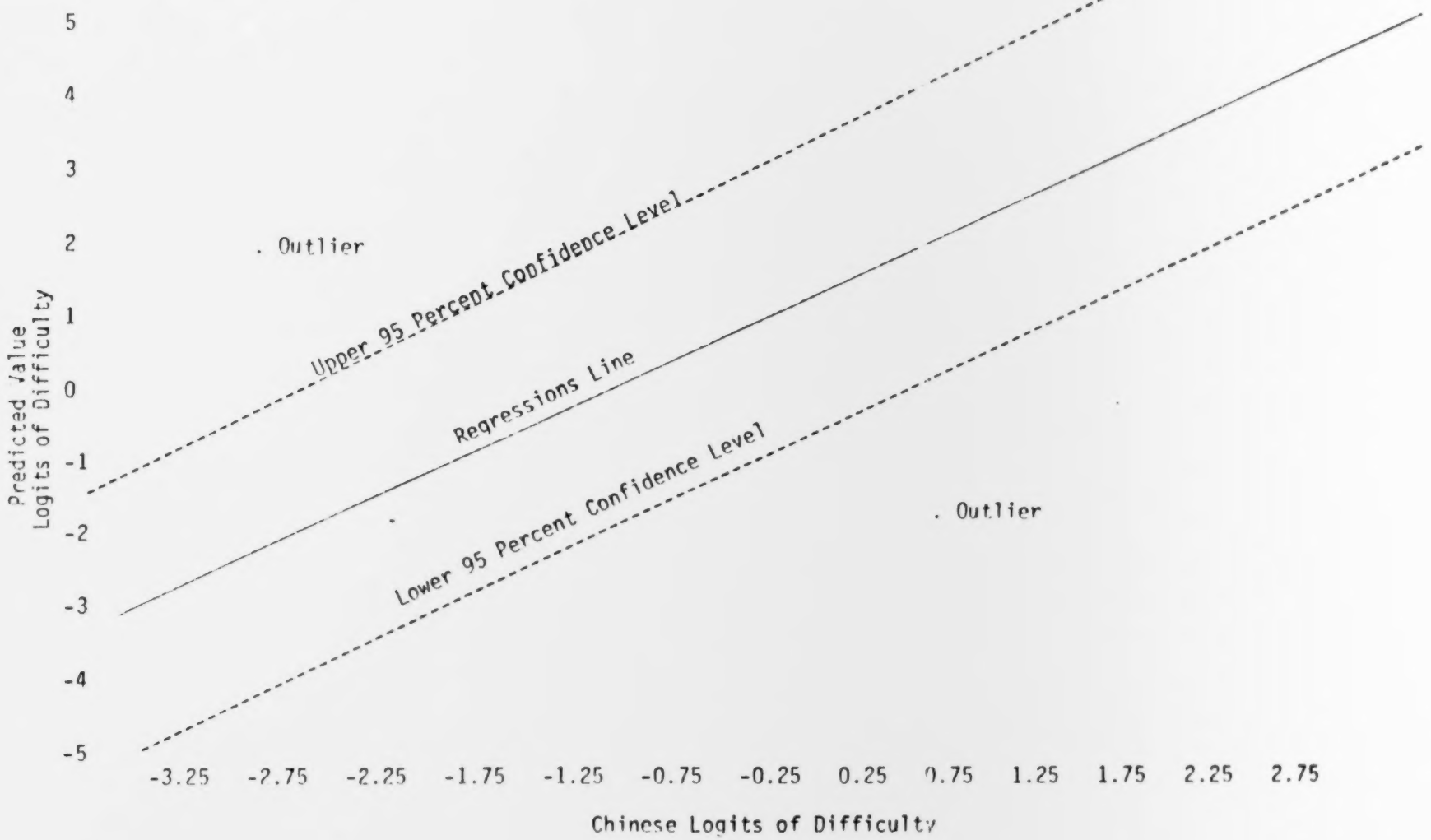

Figure 2
Distribution by Allomorph Choice

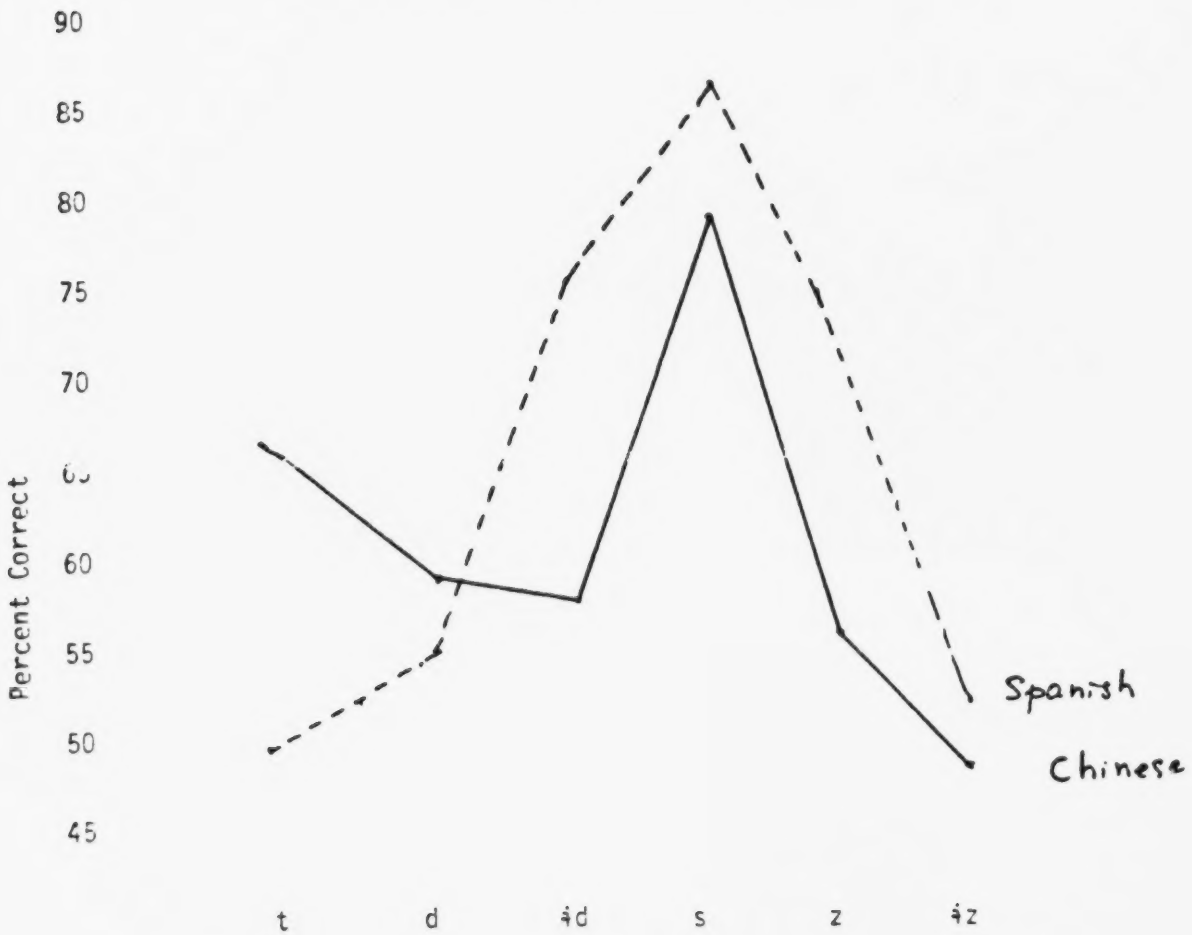

Figure 3
Distribution by manner of articulation of stem final sound

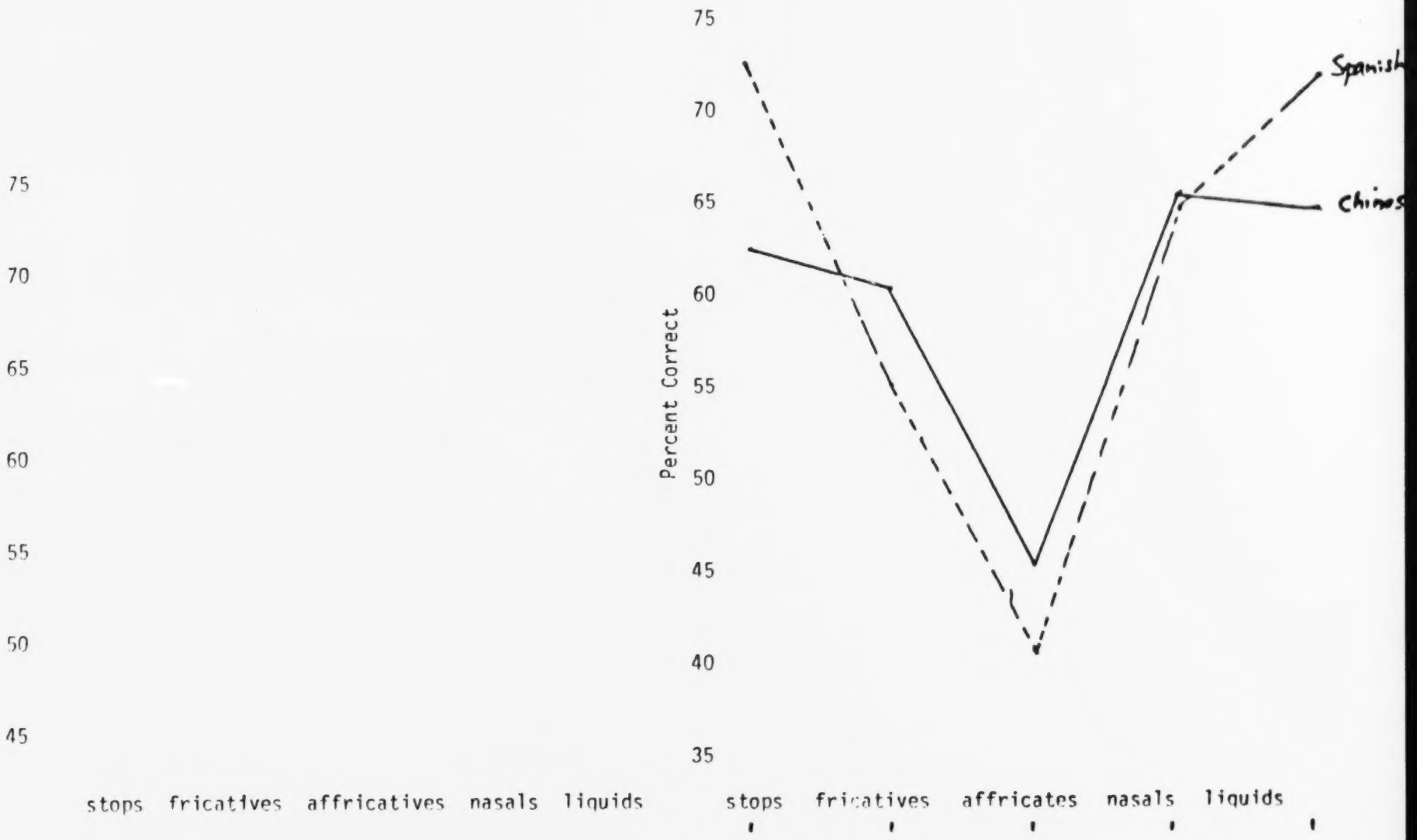

Figure 4
Distribution by place of articulation of stem final sound

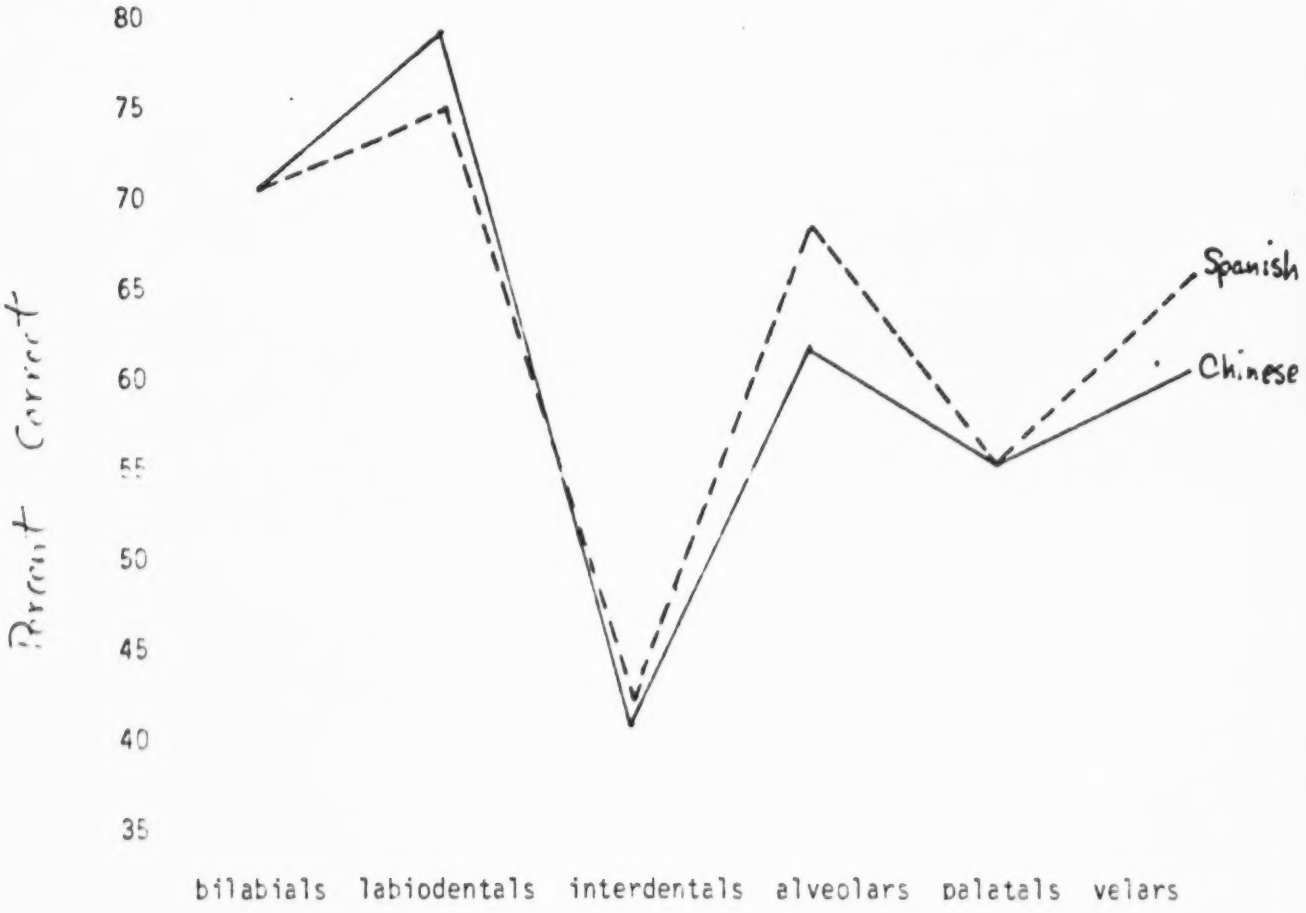

Table 1
Person Ability and Item Difficulty in Logits and the Rasch Probability of a Right Answer

Person Ability	Item Difficulty	Difference	Probability
5	0	5	.99
4	0	4	.98
3	0	3	.95
2	0	2	.88
1	0	1	.73
0	0	0	.50
0	1	-1	.27
0	2	-2	.12
0	3	-3	.05
0	4	-4	.02
0	5	-5	.01

(Wright and Stone 1979:16)

Table 2
Item Data
Chinese

Item Number	Number Correct	Item Difficulty	Point Biserial Correlation	Item Variance	Chi-Square	Logit of Difficulty	Calibration Standard Error	Outlier	Z_{ig}	Outlier
1	15	.54	.60	.25	0.003	+0.3[illegible]	.44		-0.32	
2	18	.64	.61	.23	0.064	-0.1[illegible]	.46		-0.02	
3	7	.25	.13	.19	0.019	+1.8[illegible]	.51		+0.01	
4	17	.61	.30	.24	0.006	-0.0[illegible]	.45		-0.09	
5	20	.71	.65	.21	0.112	-0.5[illegible]	.49		+0.16	
6	19	.68	.69	.22	0.059	-0.3[illegible]	.47		+0.07	
7	17	.61	.65	.24	0.006	-0.0[illegible]	.45		-0.14	
8	15	.54	.35	.25	0.043	+0.3[illegible]	.44		-0.25	
9	17	.61	.48	.24	0.006	-0.0[illegible]	.45		-0.13	
10	16	.57	.82	.24	0.246	+0.2[illegible]	.45		-0.17	
11	28	1.00	.00	.00	0.010	non-est	non-est		non-est	
12	4	.14	.38	.12	0.002	+2.6[illegible]	.63		0.00	
13	21	.75	.17	.19	0.000	-0.7[illegible]	.51		+0.14	
14	23	.82	.37	.15	0.094	-1.2[illegible]	.58		+0.35	
15	8	.29	.33	.21	1.153	+1.5[illegible]	.49		+0.05	
16	18	.64	.64	.23	0.214	-0.1[illegible]	.46		-0.02	
17	23	.82	.44	.15	0.094	-1.2[illegible]	.58		+0.39	
18	12	.43	.35	.25	0.291	+0.8[illegible]	.45		-0.18	
19	19	.68	.50	.22	0.001	-0.3[illegible]	.47		+0.02	
20	20	.71	.48	.21	0.667	-0.5[illegible]	.49		+0.10	
21	22	.79	.44	.17	0.294	-1.0[illegible]	.54		+0.27	
22	3	.11	-.14	.10	0.045	+2.9[illegible]	.71		+0.15	
23	15	.54	.33	.25	0.387	+0.3[illegible]	.44		-0.38	
24	15	.54	.68	.25	0.240	+0.3[illegible]	.44		-0.28	
25	13	.46	-.11	.25	0.979	+0.71	.44		+0.03	
26	13	.46	.21	.25	0.005	+0.71	.44		-0.02	
27	21	.75	.47	.19	0.177	-0.73	.51		+0.23	
28	14	.50	.25	.25	0.285	+0.53	.44		+0.69	yes
29	8	.29	.12	.21	1.157	+1.58	.49		+0.09	
30	20	.71	.19	.21	0.008	-0.53	.49		+0.08	
31	21	.75	.41	.19	0.015	-0.73	.51		+0.22	
32	27	.96	.05	.04	0.000	-3.19	1.19		+0.73	

Table 2 continued

Item Number	Number Correct	Item Difficulty	Point Biserial Correlation	Item Variance	Chi-Square	Logit of Difficulty	Calibration Standard Error	Outlier	Z_{ig}	Outlier
33	22	.79	.15	.17	1.742	-1.0?	.54	yes	+0.25	
34	18	.64	.32	.23	0.129	-0.1?	.46		0.00	
35	17	.61	.53	.24	0.002	-0.0?	.45		-0.14	
36	23	.82	.22	.15	0.799	-1.2?	.58		+0.35	
37	7	.25	.24	.19	0.019	+1.8?	.51		-0.03	
38	21	.75	.24	.19	0.304	-0.7?	.51		+0.17	
39	19	.68	.54	.22	0.299	-0.3?	.47		+0.02	
40	24	.86	.48	.12	0.039	-1.6?	.63		+0.48	
41	17	.61	.56	.24	0.002	-0.0?	.45		-0.14	

Table 3
Item Data
Spanish

Item Number	Number Correct	Item Difficulty	Point Biserial Correlation	Item Variance	Chi-Square	Logit of Difficulty	Calibration Standard Error	Outlier	Z_{ig}	Outlier
1	15	.54	.00	.25	0.003	+0.56	.45		+0.07	yes
2	22	.79	.45	.17	0.062	-0.83	.55		+0.29	
3	7	.25	.32	.19	0.019	+2.08	.52		-0.11	
4	18	.64	.46	.23	0.006	+0.06	.47		-0.04	
5	25	.89	.43	.10	0.107	-1.75	.73		+0.54	
6	23	.82	.30	.15	0.057	-1.07	.59		+0.36	
7	18	.64	.29	.23	0.006	+0.06	.47		-0.09	
8	13	.46	.53	.25	0.042	+0.95	.45		-0.02	
9	18	.64	.20	.23	0.006	+0.06	.47		-0.17	
10	22	.79	.28	.17	0.237	-0.83	.55		+0.27	
11	27	.96	-.04	.04	0.010	non-est	non-est		non-est	
12	5	.18	.37	.15	0.002	+2.58	.59		+0.07	
13	21	.75	.50	.19	0.000	-0.56	.52		+0.20	
14	20	.71	.38	.21	0.091	-0.32	.50		+0.08	
15	17	.61	.54	.24	1.109	+0.22	.46		-0.16	
16	24	.86	.31	.12	0.206	-1.43	.65		+0.45	
17	20	.71	.38	.21	0.091	-0.32	.50		+0.09	
18	8	.29	.44	.21	0.280	+1.84	.50		-0.03	
19	19	.68	.50	.22	0.001	-0.14	.49		+0.04	
20	13	.46	.52	.25	0.652	+0.95	.45		-0.02	
21	17	.61	.56	.24	0.283	+0.22	.46		-0.12	
22	5	.18	.28	.15	0.043	+2.58	.59		+0.09	
23	22	.79	.49	.17	0.373	-0.83	.55		+0.28	
24	11	.39	.40	.24	0.231	+1.30	.46		-0.20	
25	23	.82	.07	.15	0.945	-1.07	.59		+0.26	
26	14	.50	.57	.25	0.005	+0.76	.45		+0.06	
27	27	.96	.27	.04	0.170	-3.06	1.22		+0.70	
28	20	.71	.20	.21	0.274	-0.32	.50		+0.09	
29	4	.14	.37	.12	0.424	+2.94	.65		-0.01	
30	21	.75	.41	.19	0.007	-0.56	.52		+0.20	
31	24	.86	.32	.12	0.014	-1.43	.65		+0.44	
32	27	.96	.27	.04	0.000	-3.06	1.22		+0.70	

Table 3 continued

Item Number	Number Correct	Item Difficulty	Point Biserial Correlation	Item Variance	Chi-Square	Logit of Difficulty	Calibration Standard Error	Outlier	Z_{ig}	Outlier
33	11	.39	.46	.24	1.677	+1.30	.46		-0.23	
34	23	.82	.24	.15	0.124	-1.07	.59		+0.36	
35	17	.61	.31	.24	0.002	+0.27	.46		-0.26	
36	15	.54	.51	.25	0.769	+0.56	.45		+0.17	
37	7	.25	.72	.19	0.019	+2.08	.52		-0.17	
38	16	.57	.61	.25	0.293	+0.47	.46		-0.20	
39	26	.93	-.05	.06	0.288	-2.35	.88		+0.58	
40	22	.79	.52	.17	0.038	-0.83	.55		+0.30	
41	17	.61	.35	.24	0.002	+0.27	.46		-0.21	